CUNY SPS DATA 621 - CTG5 - HW4

Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh

April 24th, 2019

Contents

Table 1: Data Dictionary

VARIABLE	DEFINITION	TYPE
TARGET_FLAG	car crash = 1, no car crash = 0	binary categorical response
TARGET_AMT	car crash cost = >0, no car crash = 0	continuous numerical response
AGE	driver's age - very young/old tend to be risky	continuous numerical predictor
BLUEBOOK	$ value of vehicle	continuous numerical predictor
CAR_AGE	age of vehicle	continuous numerical predictor
CAR_TYPE	type of car (6types)	categorical predictor
CAR_USE	usage of car (commercial/private)	binary categorical predictor
CLM_FREQ	number of claims past 5 years	discrete numerical predictor
EDUCATION	max education level (5types)	categorical predictor
HOMEKIDS	number of children at home	discrete numerical predictor
HOME_VAL	$ home value - home owners tend to drive more responsibly	continuous numerical predictor
INCOME	$ income - rich people tend to get into fewer crashes	continuous numerical predictor
JOB	job category (8types, 1missing) - white collar tend to be safer	categorical predictor
KIDSDRIV	number of driving children - teenagers more likely to crash	discrete numerical predictor
MSTATUS	maritial status - married people drive more safely	catogerical predictor
MVR_PTS	number of traffic tickets	continuous numerical predictor
OLDCLAIM	$ total claims in the past 5 years	continuous numerical predictor
PARENT1	single parent	binary categorical predictor
RED_CAR	a red car	binary categorical predictor
REVOKED	license revoked (past 7 years) - more risky driver	binary categorical predictor
SEX	gender - woman may have less crashes than man	binary categorical predictor
TIF	time in force - number of years being customer	continuous numerical predictor
TRAVTIME	distance to work	continuous numerical predictor
URBANCITY	urban/rural	binary categorical predictor
YOJ	years on job - the longer they stay more safe	continuous numerical predictor

1 DATA EXPLORATION

In this assignment we explore, analyze and model a dataset containing 8,161 observations with 25 variables each representing a customer at an auto insurance company. Two of the 25 features are target varibles and 23 are predictors. One of the target variables, `TARGET_FLAG`, is a binary categorical variable where a value of 1 indicates that the customer has made a claim related to a car crash and a value of 0 indicates they have not. The other target variable, `TARGET_AMT`, is a continuous numerical variable representing the payout amount of a claim, if any. Of the remaining 23 predictor variables, 13 are categorical and 10 are numerical.

Using this data, we will compose and evaluate several types of models with the following objectives: - Logistic classification models that aim to predict the probability that a person crashes their car - Multiple linear regression models that aim to predict the amount of money it will cost if the person does crash their car

The intended use case for these models is actuarial in nature: specifically, to calculate insurance rates commensurate with policyholders' (or policy applicants') potential risk levels, based on attributes such as income, age, distance to work, tenure as customers, etc.

Inspection of the target variables reveals that where `TARGET_FLAG` has values of 0 (i.e. no claim) `TARGET_AMT` also has values of 0 (i.e. no payout), which is logically consistent.

1.1 Summary Statistics

We summarize contintuous and categorical variables separately.

Table 2: (#tab:t2.1)Summary statistics

	n	min	mean	median	max	sd
TARGET_AMT	8161	0	1504.3	0	107586	4704.0
AGE	8155	16	44.8	45	81	8.6
YOJ	7707	0	10.5	11	23	4.1
INCOME	7716	0	61898.1	54028	367030	47572.7
HOME_VAL	7697	0	154867.3	161160	885282	129123.8
TRAVTIME	8161	5	33.5	33	142	15.9
BLUEBOOK	8161	1500	15709.9	14440	69740	8419.7
TIF	8161	1	5.3	4	25	4.2
OLDCLAIM	8161	0	4037.1	0	57037	8777.1
MVR_PTS	8161	0	1.7	1	13	2.1
CAR_AGE	7651	0	8.3	8	28	5.7

Table 3: (#tab:t2.2)Summary statistics for Categorical Variables

	EDUCATION	JOB	CAR_TYPE	KIDSDRIV	HOMEKIDS	CLM_FREQ
	<High School :1203	z_Blue Collar:1825	Minivan :2145	0:7180	0:5289	0:5009
	Bachelors :2242	Clerical :1271	Panel Truck: 676	1: 636	1: 902	1: 997
	Masters :1658	Professional :1117	Pickup :1389	2: 279	2:1118	2:1171
	PhD : 728	Manager : 988	Sports Car : 907	3: 62	3: 674	3: 776
	z_High School:2330	Lawyer : 835	Van : 750	4: 4	4: 164	4: 190
	NA	Student : 712	SUV :2294	NA	5: 14	5: 18
	NA	(Other) :1413	NA	NA	NA	NA

`EDUCATION`, `JOB`, `CAR_TYPE`, `KIDSDRIV`, `HOMEKIDS`, AND `CLM_FREQ` each comprise multiple categories. On the other hand, `PARENT1`, `SEX`, `MSTATUS`, `CAR_USE`, RED_CAR,REVOKED`,` `URBANICITY` are all binaries.

Examining the dispersion of claims betweent variables, it looks like likelihoods are higher for drivers who are male, urban, blue collar, unmarried, or parents; as well as for those with commerical vehicles or a revoked license.

Table 4: (#tab:t2.2)Summary statistics for Binary Categorical Variables

	PARENT1	SEX	MSTATUS	CAR_USE	RED_CAR	REVOKED	URBANICITY
	No :7084	M:3786	Yes:4894	Commercial:3029	no :5783	No :7161	Urban:6492
	Yes:1077	F:4375	No :3267	Private :5132	yes:2378	Yes:1000	Rural:1669

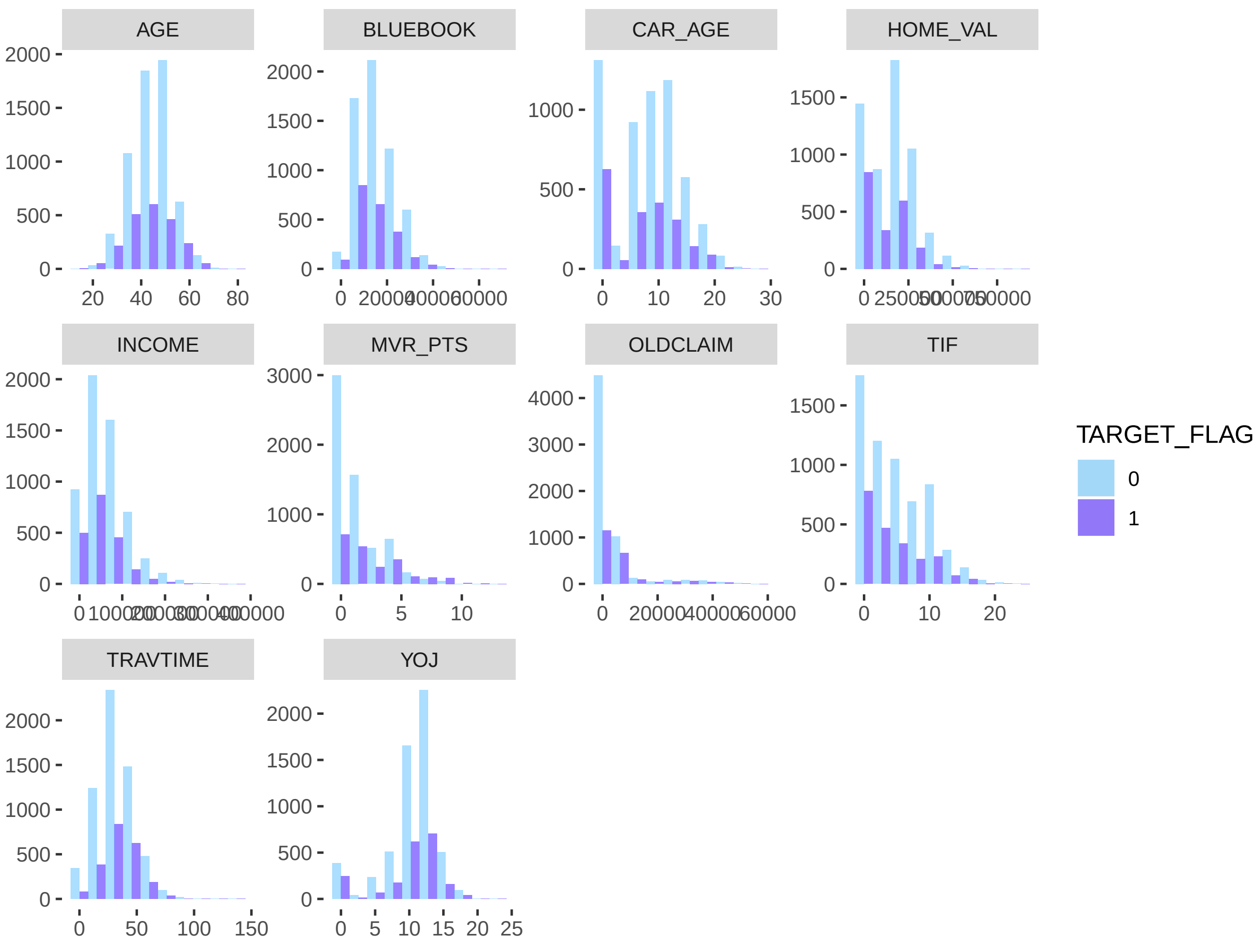


Figure 1: Numeric Data Distributions as a Function of TARGET_FLAG

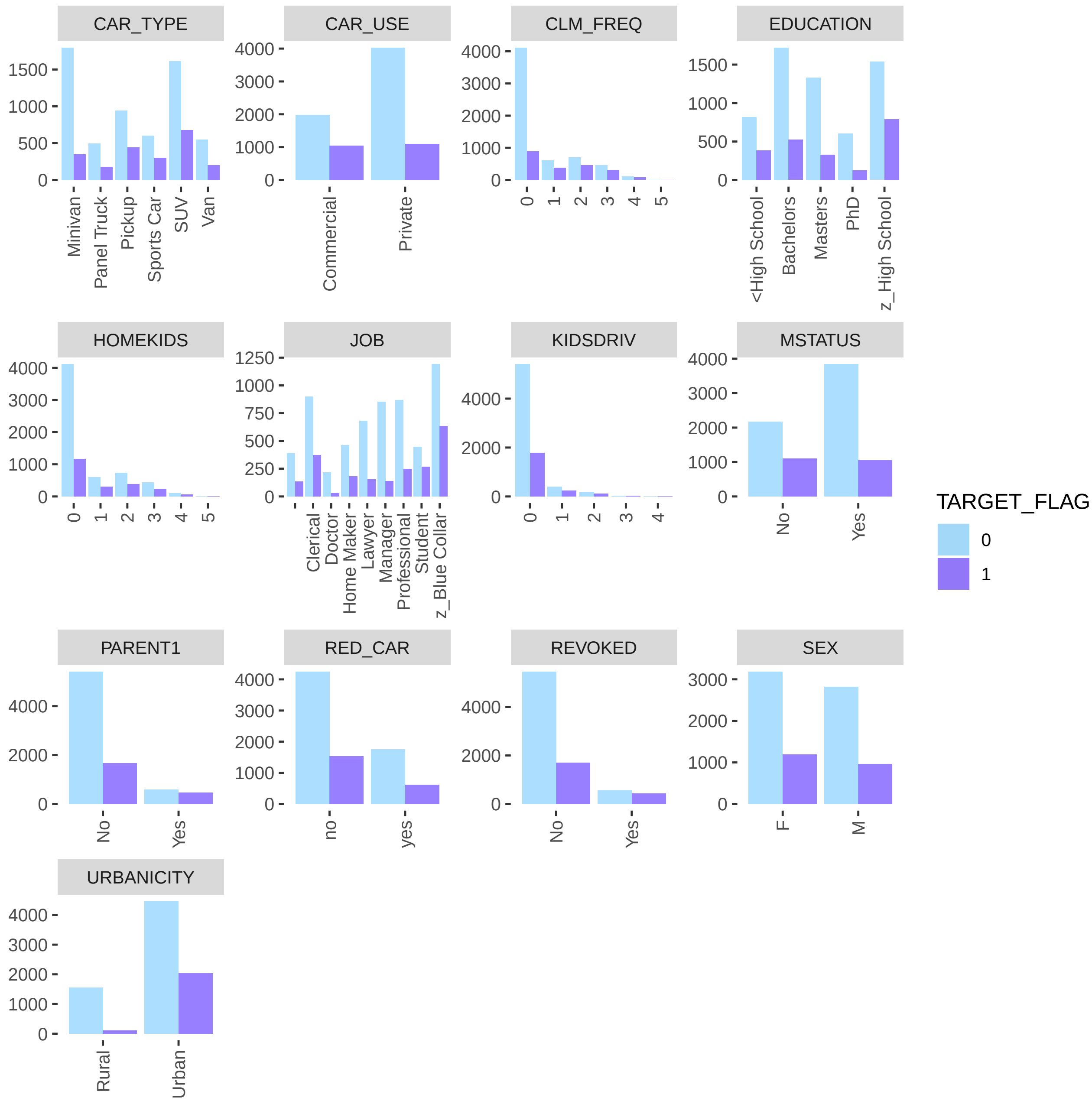


Figure 2: Categorical Data Distributions as a Function of TARGET_FLAG

The scale of continuous variables' distributions are considerably different and difficult to visualize together.

Scaling the distribution of based on the standard deviation reveals that outliers are very abundant for certain continuous variables, particularly `OLDCLAIM`, `INCOME`, `TRAV_TIME`, `BLUEBOOK`, and to a lesser extent `HOME_VAL` and `TIF`.

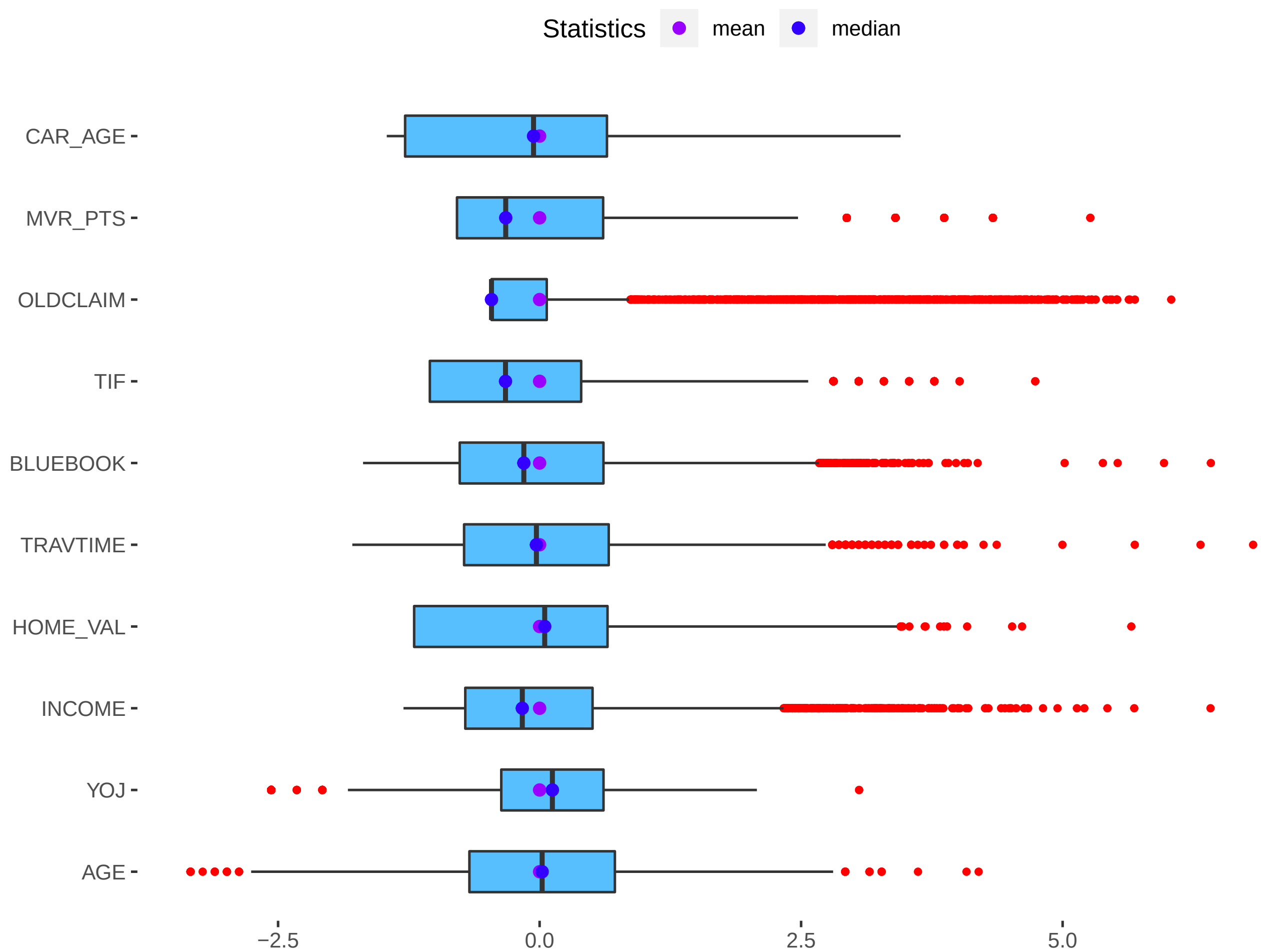


Figure 3: Scaled Boxplots

[JO: WEREN'T WE GO TO TOSS THE NEXT CHART?]

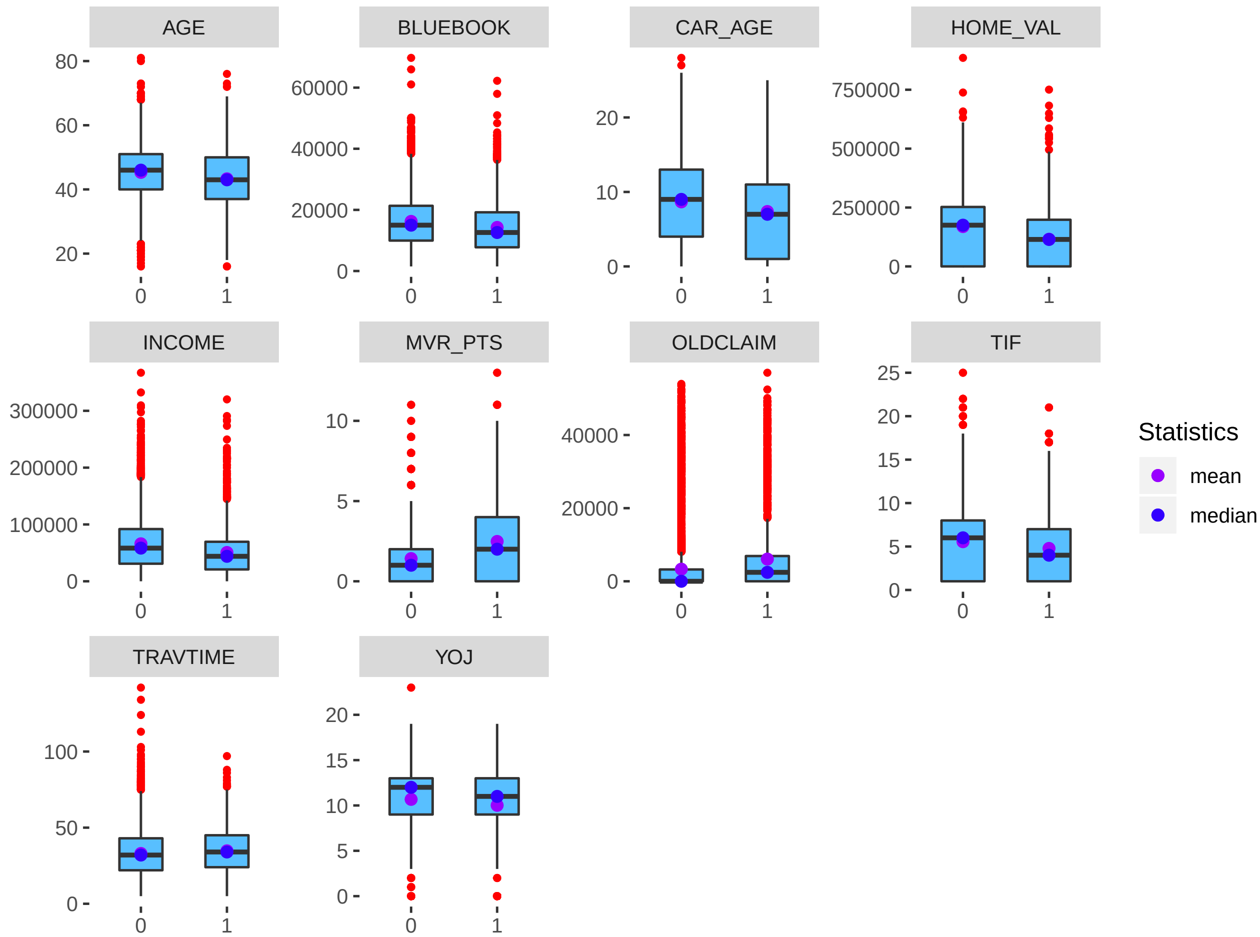


Figure 4: Linear relationship between each numeric predictor and the target

1.2 Linearity

[JO: DON'T THINK THE LOG TRANSFORM DISAMBIGUATES LINEAR RELATIONSHIPS - IF TEAM ALIGNED, WE CAN TRY TO ADD REGRESSION EQUATIONS TO THE FACETS: https://community.rstudio.com/t/annotate-ggplot2-with-regression-equation-and-r-squared/6112/6]

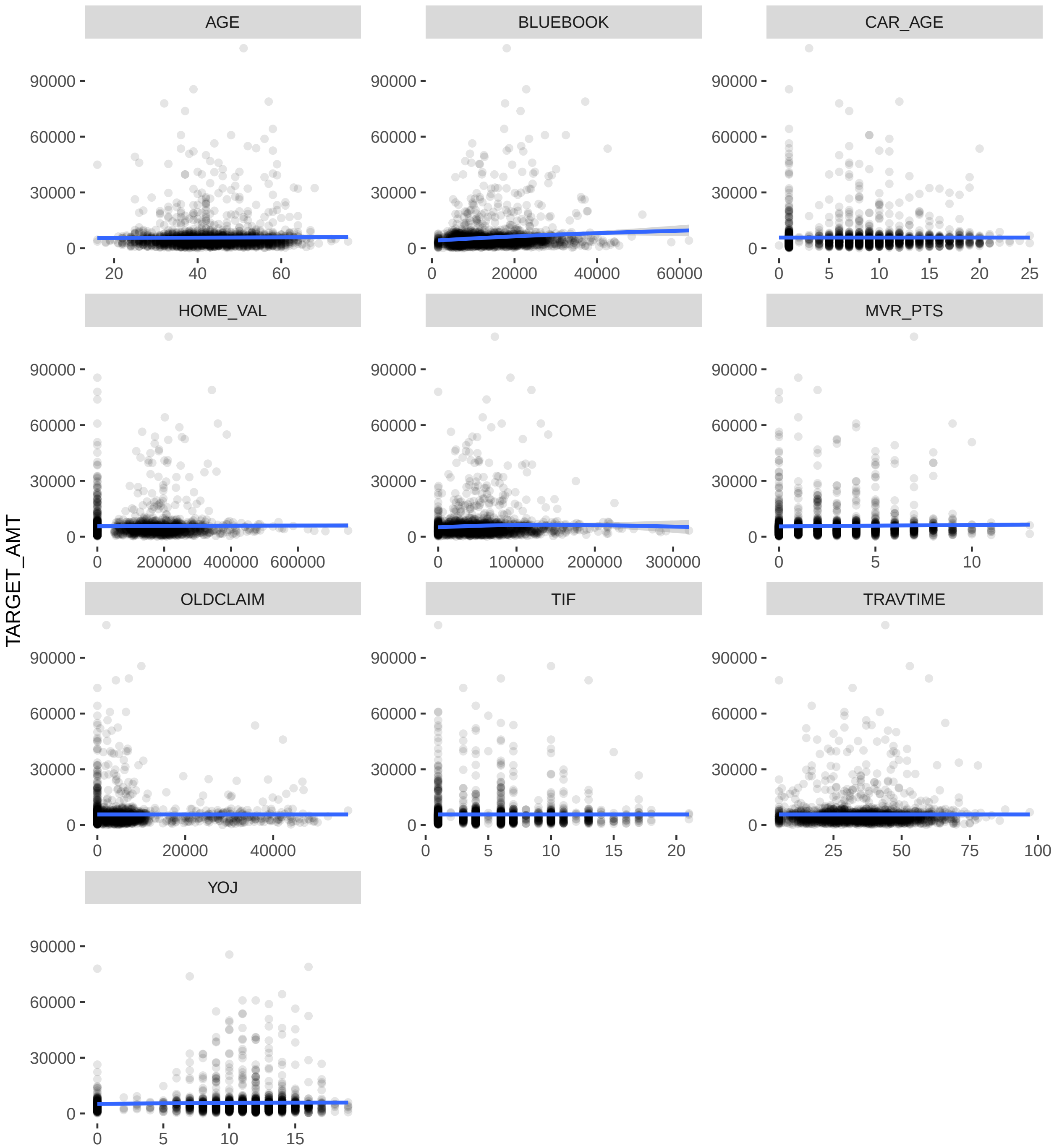


Figure 5: Scatter plot between numeric predictors and the TARGET_AMT

[JO: DON'T THINK THE LOG TRANSFORM HAS HELPED - THINK WE SHOULD ADD]

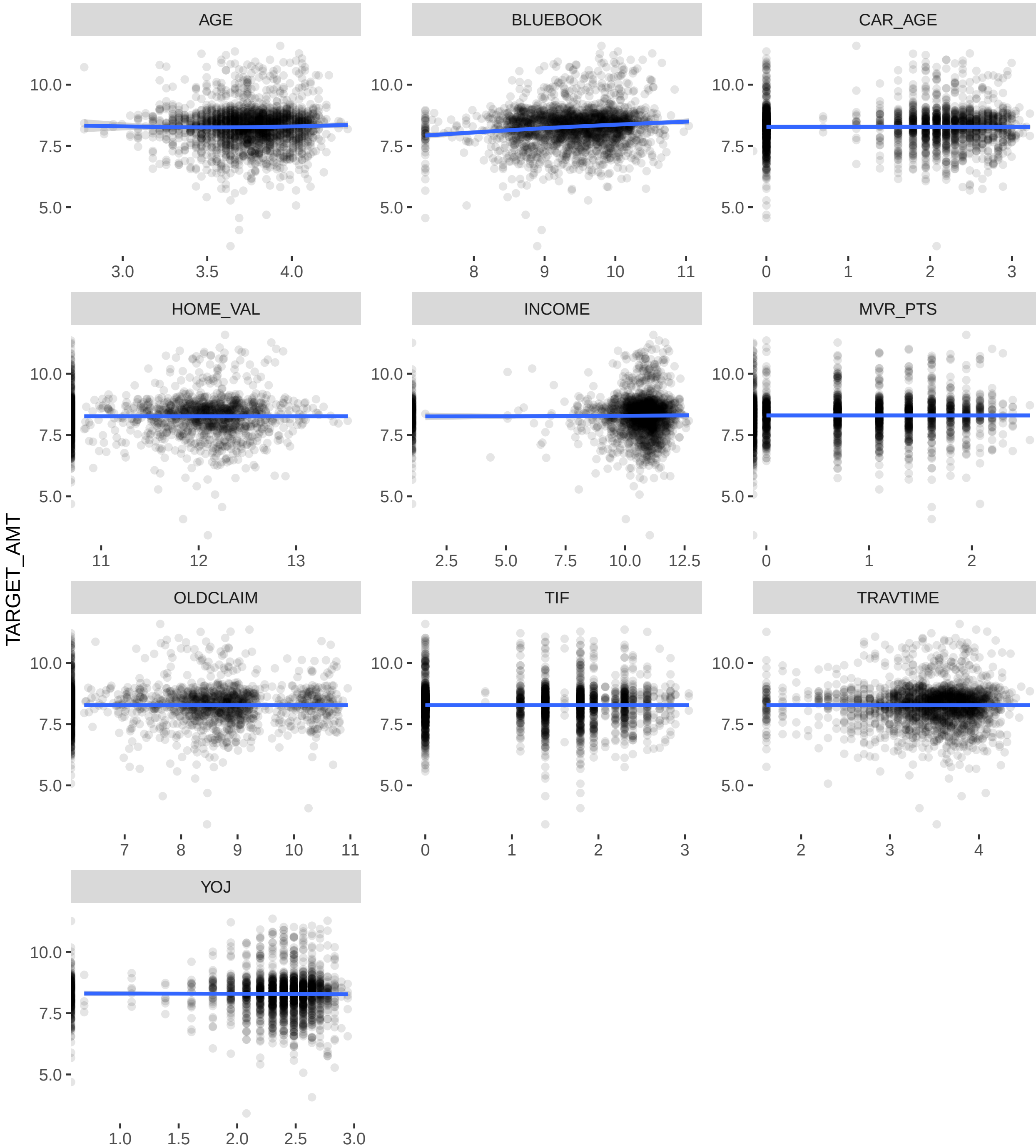


Figure 6: Scatter plot between log transformed numeric predictors and the log transformed TARGET_AMT

1.3 Missing Data

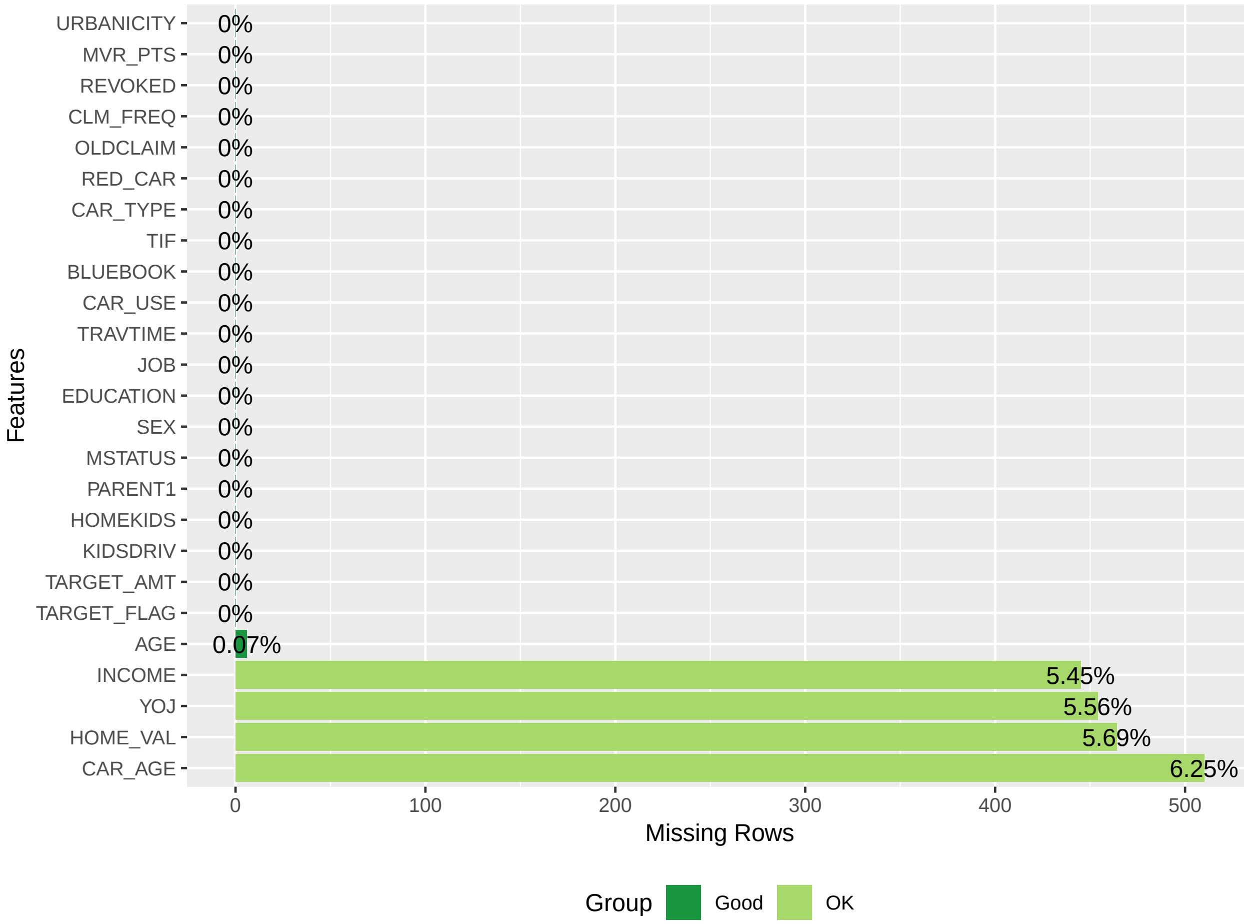


Figure 7: Missing data

A number of variables are missing observations: `AGE`, `INCOME`, `YOJ`, `HOME_VAL`, `CAR_AGE`. For `AGE`, the number is not consequential, but the others range between 5% and 6% of total.

[JO: WHERE DID WE NET OUT ON CHECKING WHETHER THE SAME OBSERVATIONS ARE MISSING THE ABOVE VARIABLES AND WHETHER THEY HAVE ANYTHING IN COMMON I.E. THEY'RE ALL MARRIED, HIGH EARNERS, HOMEOWNERS?]

We will impute missing values when preparing the data for modelling.

2 DATA PREPARATION

2.1 Variable Descriptions

2.1.0.1 KIDSDRIV

`KIDSDRIV` is a categorical predictor with values ranging from 0 to 4. It shows heavy skew, with most cars having no kid drivers (value of 0). Judging from the distribution, it appears that having kid driver results in higher probability of making a claim.

2.1.0.2 AGE

`AGE` presents driver's age and shows a normal distribution centered around 45 years. Looking at the boxplot of age, there does not appear to be a difference in the distribution between whether a claim is made or not. Accordingly, this `AGE` may not be helpful in determining the probability of making a claim.

2.1.0.3 HOMEKIDS

`HOMEKIDS` is a predictor describing number of children at home ranging from 0 to 5.

2.1.0.4 YOJ

`YOJ` is a predictor describing years on job. People who stay at a job for a longer time are believed to be safer drivers. Apart from those who are unemployed (values of 0), `YOJ` seems to show a normal distribution.

2.1.0.5 INCOME

`INCOME` is a heavily skewed predictor variable, suggesting thatoutliers should be treated for modelling.

2.1.0.6 HOME_VAL

`HOME_VAL` is a home value predictor variable. In theory, home owners tend to drive more responsibly. The difference between owners and renters (values of 0) is visible in the graph.

2.1.0.7 TRAVTIME

`TRAVTIME` is a predictor variable describing the distance to work. Long drives to work would suggest greater risk of an accident and claim. However, the graph shows a fairly normal distribution, such that this variable may not be helpful in determining the probability of making a claim.

2.1.0.8 BLUEBOOK

`BLUEBOOK` is a predictor variable describing the value of the car. The'boxplot demonstrates that the lower value of the car, the higher chances of making a claim. It is conceivable that higher-priced cars are driven more carefully.

2.1.0.9 TIF

`TIF` describes how long the customer has been with the insurance company. Plots reveal that the longer the tenure of a policyholder, the lower the likelihood of a claim - i.e. safe drivers tend to remain so.

2.1.0.10 OLDCLAIM

`OLDCLAIM` is a predictor describing the value of claims made in the past 5 years. It is very heavily skewed as most policyholders do not make claims.

2.1.0.11 CLM_FREQ

`CLM_FREQ` is a predictor that describing the frequency of claims in the past 5 years. It suggestst that those who have made a claim in the past 5 years are more likely to make another claim.

2.1.0.12 MVR_PTS

`MVR_PTS` is a predictor that describes motor vehicle record points. The rationale is that more traffic tickets suggests less safe driving and a higher likelihood of claims. It appears to be a highly significant variable as seen in boxplots.

2.1.0.13 CAR_AGE

`CAR_AGE` describes the age of the policyholder's vehicle. One value is -3, which must be an error - this is corrected to 0.

2.1.0.14 PARENT1

`PARENT1` indicates whether a policyholder is a single parent. This variable has been factorized and relabeled as `NumParents` to describe the number of parents.

2.1.0.15 SEX

`SEX` describes the gender of the driver. This variable has been factorized and relabeled as 'MALE, for which males receive a value of 1 and females a value of as 0. It does not appear to be significant variable in the box plot.

2.1.0.16 MSTATUS

`MSTATUS` describes the marital status of the policyholder. The rationale is that married people drive more safely. This variable has been factorized and relabeled as `Single`, for which married policgholders receive a value of 0 and unmarried a value o 1.

2.1.0.17 EDUCATION

`EDUCATION` describes the education level of the driver. This variable is factorized. It may be correlated with INCOME.

2.1.0.18 JOB

`JOB` describes the type of job the driver has. This variable is factorized. It may be correlated with INCOME. In theory policyholders with white collar jobs tend to drive more safely.

2.1.0.19 CAR_TYPE

`CAR_TYPE` describes type of car. This variable is factorized.

2.1.0.20 CAR_USE

CAR_USE describes how the vehicle is used. Commercial vehicles are driven more and may have an elevated probability of accidents and claims. This variable is factorized and relabeled as Commercial, for which a value of 0 means private use and a value of 1 means commercial use.

2.1.0.21 RED_CAR

RED_CAR indicates whether the color of the vehicle is red. Red vehicles, especially sports cars, are associated with riskier driving and likelihood of claims. This variable is factorized.

2.1.0.22 REVOKED

REVOKED describes whether a policyholders license has been revoked in the past 7 years. License revocatiom is associated with riskier driving. This variable is factorized. The boxplot reveals that policyholders who previously lost their license are more likely to file claims.

2.1.0.23 URBANICITY

URBANICITY describes whether driver lives in Urban area or Rural area. This variable has been factorized and relabeled as URBAN, for which a value of 0 means rural and a value of 1 means urban.

2.2 Missing values

[JO: THINK WE SHOULD DESCRIBE PURPOSE OF/ NEED FOR VALUE IMPUTATION. MICE IMPUTATION ASSUMES 'MISSING AT RANDOM' (MAR), SO THINK WE'LL NEED TO ESTABLISH THAT THIS IS THE CASE.][JO: WHAT'S THE DIFFERENCE BETWEEN THE M AND MAXIT VALUES (1 FOR AGE, 2 FOR OTHERS?)]

To deal with missing data values for the variables INCOME, YOJ, HOME_VAL, and CAR_AGE - and to a lesser extentAGE' - we leveraged the MICE (Multivariate Imputation By Chained Equations) package. The package creates multiple imputations (replacement values) for multivariate missing data using a a method based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model. The method can impute mixes of continuous, binary, unordered categorical and ordered categorical, and continuous two-level data; and it can maintain consistency between imputations by means of passive imputation. We inspected the quality of imputed values using multiple diagnostic plots.

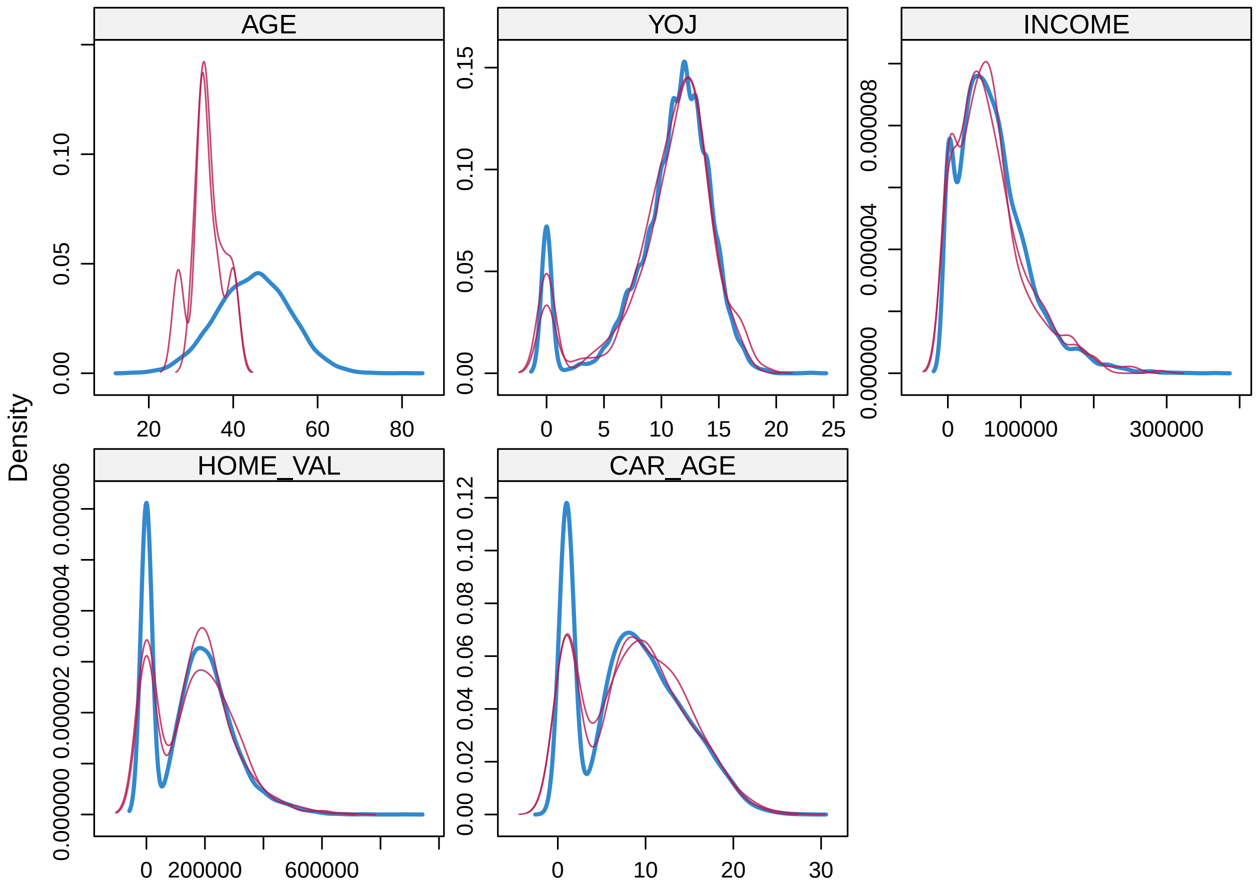
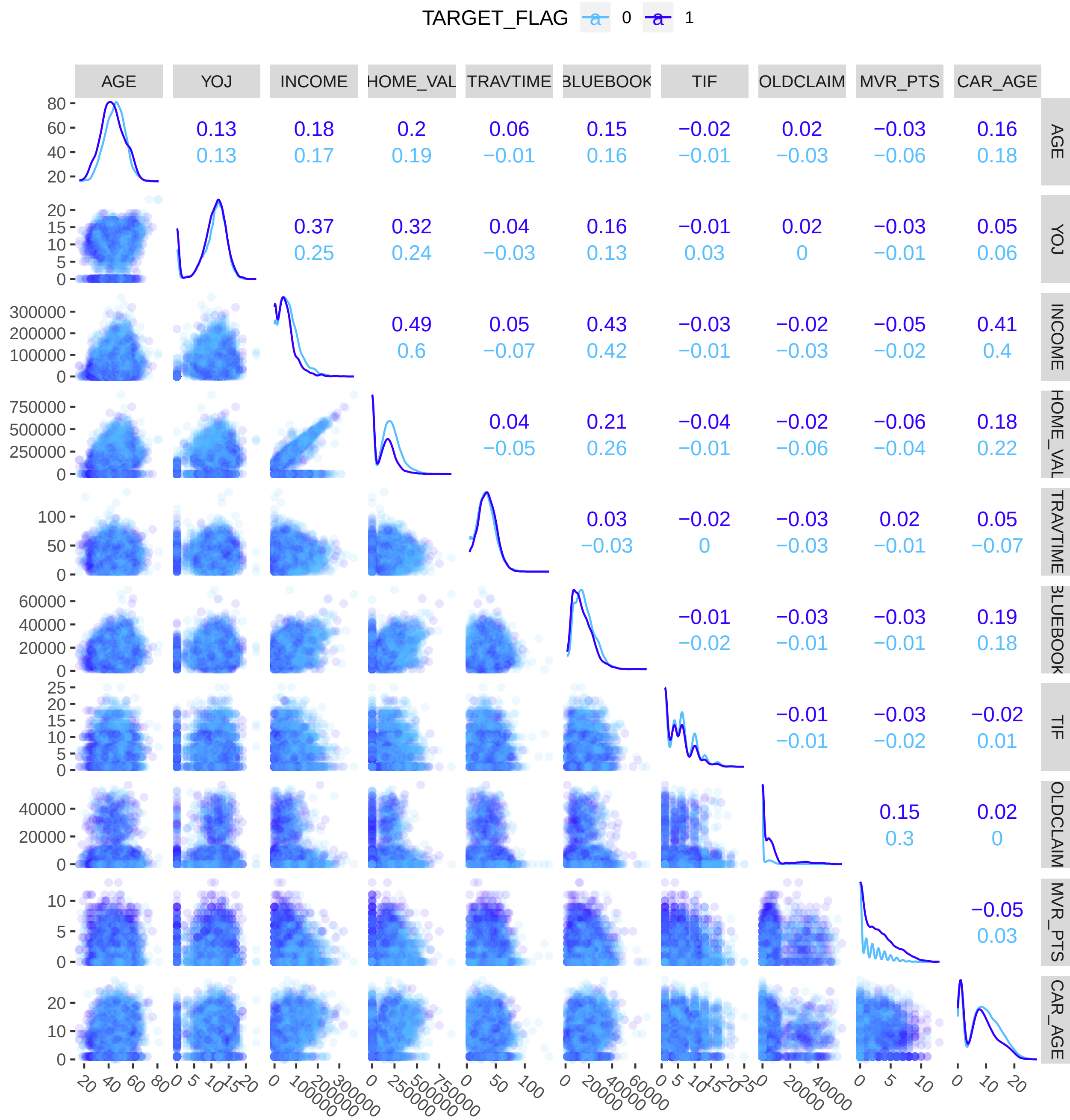
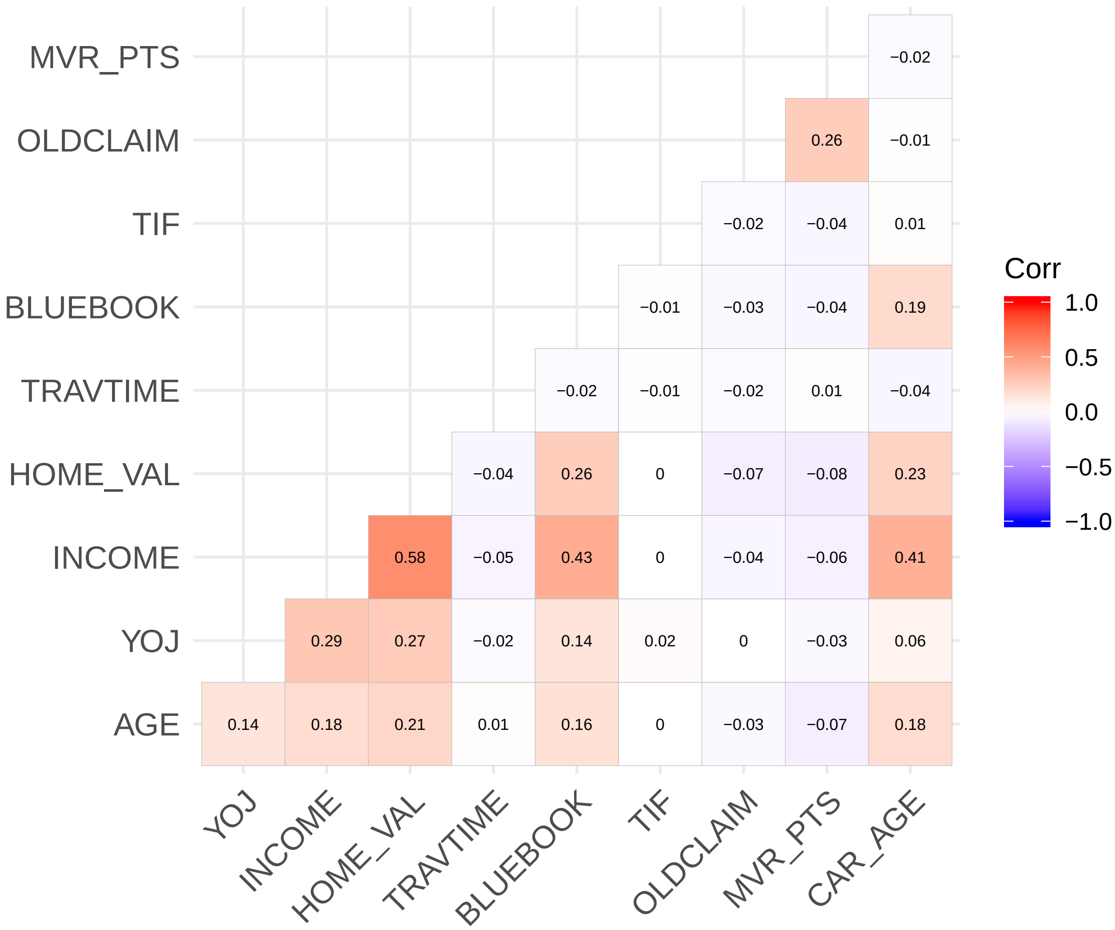


Figure 8: Difference between original and imputed data

The blue and red lines represent the distribution originally known and imputed values respectively. With the exception of `AGE`, the distribution of the imputed values accords with the distribution of where values are not missing. For those four variables we will use the imputed values. For `AGE` (only .07 or XX number of cases) we will impute `AGE` separately using the median imputation.



Unsurpisingly, higher levels of `INCOME` are found with higher values of `YOJ`; this also means more income is disposable, which shows correlation with HOME_VAL and BLUEBOOK. Additionally, `MVR_PTS` shows a relationship with `OLDCLAIMS`.


MVR_PTS
OLDCLAIM
TIF
BLUEBOOK
TRAVTIME
HOME_VAL
INCOME
YOJ
AGE
YOJ
INCOME
HOME_VAL
TRAVTIME
BLUEBOOK
TIF
OLDCLAIM
MVR_PTS
CAR_AGE
Corr
1.0
0.5
0.0
−0.5
−1.0
−0.02
0.26
−0.01
−0.02
−0.04
0.01
−0.01
−0.03
−0.04
0.19
−0.02
−0.01
−0.02
0.01
−0.04
−0.04
0.26
0
−0.07
−0.08
0.23
0.58
−0.05
0.43
0
−0.04
−0.06
0.41
0.29
0.27
−0.02
0.14
0.02
0
−0.03
0.06
0.14
0.18
0.21
0.01
0.16
0
−0.03
−0.07
0.18

3 BUILD MODELS

3.1 Classification Models: Model 1,2,3,4

The first four models take categorical variables as inputs and interpret their contributions to predicting the likelihood of a claim [JO: CONFIRM JUST 'for new customers'?]. We use `drop` and `MASS:stepAIC` functions to judge which variables to remove, evaluating AIC statistics as we go.

3.1.1 MODEL 1

___ TARGET_FLAG ~ NumParents+ Male+ EDUCATION+ JOB+ CAR_TYPE+ RED_CAR+ REVOKED+ Urban+ Single+ Commercial ___

For an easily interpretable model aimed at predict TARGET_FLAG, we restrict inputs for Model 1 to categorical variables alone. The AIC metric suggests that the `RED_CAR` variable can be removed.

3.1.2 MODEL 2

___TARGET_FLAG ~ KIDSDRIV+ AGE+ HOMEKIDS + YOJ+INCOME+HOME_VAL+ TRAVTIME+ BLUEBOOK+ TIF+OLDCLAIM+ CLM_FREQ+ MVR_PTS+ CAR_AGE + PARENT1+ SEX+ EDUCATION+ JOB+ CAR_TYPE+ REVOKED+ URBANICITY+ MSTATUS+ CAR_USE

As suggested by the Model 1 finding, Model 2 excludes the `RED_CAR` variable. AIC metrics suggest removing `AGE`, `CAR_AGE` and `SEX`.

3.1.3 MODEL 3

TARGET_FLAG ~ KIDSDRIV+ HOMEKIDS + YOJ+INCOME+HOME_VAL+ TRAVTIME+ BLUEBOOK+ TIF+OLDCLAIM+ CLM_FREQ+ MVR_PTS+ PARENT1+ EDUCATION+ JOB+ CAR_TYPE+ REVOKED+ URBANICITY+ MSTATUS+ CAR_USE

Building on model 2, model 3 excludes `AGE`, `CAR_AGE` and `SEX`.

[RJ: Anybody wants to create classification evaluation table in the select model section for those? − might want to roll back for the data cleaning I had done previously as we need the factorized values to create that matrices.]

3.1.4 Model 4 - Binary logistic model

Model 4 incorporates all explanatory variables plus log transformations of skewed variables: `INCOME`, `TRAVTIME`, `BLUEBOOK`, `OLDCLAIM`, and `AGE` in a binary logistic model refined through backward elimination. *Lidiia: I am not sure why, but there is a formatting issue. The summary for this and all other models below are at the end of the pdf.*

Observations	8161
Dependent variable	TARGET_FLAG
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(35)$	2140.47
Pseudo-R² (Cragg-Uhler)	0.34
Pseudo-R² (McFadden)	0.23
AIC	7349.49
BIC	7601.75

	Est.	S.E.	z val.	p	VIF
(Intercept)	2.04	0.79	2.57	0.01	NA
KIDSDRIV1	0.58	0.11	5.54	0.00	1.15
KIDSDRIV2	0.82	0.15	5.40	0.00	1.15
KIDSDRIV3	0.98	0.30	3.25	0.00	1.15
KIDSDRIV4	1.38	1.12	1.23	0.22	1.15
log(AGE)	-0.31	0.16	-2.00	0.05	1.26
YOJ	0.01	0.01	1.30	0.19	2.37
log(INCOME + 0.00000000000001)	-0.02	0.00	-4.07	0.00	3.23
HOME_VAL	-0.00	0.00	-5.25	0.00	1.75
log(TRAVTIME)	0.41	0.05	7.94	0.00	1.03
log(BLUEBOOK)	-0.32	0.06	-5.87	0.00	1.48
TIF	-0.05	0.01	-7.34	0.00	1.01
log(OLDCLAIM + 0.00000000000001)	0.01	0.00	6.24	0.00	1.26
MVR_PTS	0.10	0.01	7.09	0.00	1.24
PARENT1Yes	0.37	0.10	3.66	0.00	1.64
EDUCATIONBachelors	-0.41	0.11	-3.75	0.00	7.47
EDUCATIONMasters	-0.33	0.16	-2.04	0.04	7.47
EDUCATIONPhD	-0.30	0.19	-1.53	0.13	7.47
EDUCATIONz_High School	0.03	0.09	0.35	0.73	7.47
JOBClerical	0.49	0.19	2.53	0.01	26.60
JOBDoctor	-0.39	0.27	-1.48	0.14	26.60
JOBHome Maker	0.15	0.21	0.69	0.49	26.60
JOBLawyer	0.16	0.17	0.96	0.34	26.60
JOBManager	-0.51	0.17	-2.99	0.00	26.60
JOBProfessional	0.22	0.18	1.26	0.21	26.60
JOBStudent	0.12	0.22	0.55	0.58	26.60
JOBz_Blue Collar	0.39	0.19	2.10	0.04	26.60
CAR_TYPEPanel Truck	0.54	0.14	3.76	0.00	2.33
CAR_TYPEPickup	0.58	0.10	5.80	0.00	2.33
CAR_TYPESports Car	0.96	0.11	8.85	0.00	2.33
CAR_TYPEVan	0.65	0.12	5.32	0.00	2.33
CAR_TYPESUV	0.74	0.09	8.57	0.00	2.33
REVOKEDYes	0.71	0.08	8.82	0.00	1.01
URBANICITYRural	-2.35	0.11	-20.86	0.00	1.14
MSTATUSNo	0.46	0.08	5.62	0.00	1.96
CAR_USEPrivate	-0.75	0.09	-8.18	0.00	2.46

Standard errors: MLE

3.2 Regression Model: Model 5,6

The next two models are multiple linear regression models aimed at predicting the value of claims based on different approaches, including constraining the cases based on `TARGET_FLAG` (i.e. based on whether or not a claim was filed) and different approaches to selecting explanatory variables.

3.2.1 Model 5 - Multiple linear regression model

Model 5 is a multiple linear regression model built only on cases with claims where `TARGET_AMT` is greater than 0. The model is refined using stepwise elimination. [JO: WE SHOULD ADD INTERPRETATION OF T- AND P-VALS]..

Observations	2153
Dependent variable	TARGET_AMT
Type	OLS linear regression

F(51,2101)	1.50
R^2	0.04
Adj. R^2	0.01

3.2.2 Model 6 - Multiple linear regression model

Model 6 includes is a multiple linear regression model built on all cases - in other words it relaxes the constraint that a claim was filed, and so includes `TARGET_AMT` values of 0. Any predicted value less than $100 will be considered 0. [JO: WHY IS THIS?] Forward elimination is used to refine variable selection.

4 SELECT MODELS

5 Appendix

The appendix is available as script.R file in `project4_insurance` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining

	Est.	S.E.	t val.	p
(Intercept)	-2748.39	21120.36	-0.13	0.90
KIDSDRIV1	379.71	639.54	0.59	0.55
KIDSDRIV2	-236.42	857.06	-0.28	0.78
KIDSDRIV3	-733.23	1504.06	-0.49	0.63
KIDSDRIV4	-674.97	6818.08	-0.10	0.92
log(AGE)	-6862.94	4904.16	-1.40	0.16
AGE	178.06	118.18	1.51	0.13
HOMEKIDS1	28.49	695.60	0.04	0.97
HOMEKIDS2	826.71	681.45	1.21	0.23
HOMEKIDS3	309.98	769.38	0.40	0.69
HOMEKIDS4	521.92	1186.80	0.44	0.66
HOMEKIDS5	205.53	3946.35	0.05	0.96
YOJ	46.09	63.85	0.72	0.47
log(INCOME + 0.00000000000001)	-9.07	23.76	-0.38	0.70
INCOME	-0.01	0.01	-1.26	0.21
log(TRAVTIME)	-313.68	838.43	-0.37	0.71
TRAVTIME	11.48	31.12	0.37	0.71
log(BLUEBOOK)	1015.07	686.49	1.48	0.14
BLUEBOOK	0.04	0.06	0.66	0.51
TIF	-21.76	42.91	-0.51	0.61
log(OLDCLAIM + 0.00000000000001)	-534.94	483.14	-1.11	0.27
OLDCLAIM	0.07	0.04	1.61	0.11
CLM_FREQ1	21333.56	19437.05	1.10	0.27
CLM_FREQ2	20886.98	19434.87	1.07	0.28
CLM_FREQ3	21314.36	19439.45	1.10	0.27
CLM_FREQ4	20870.82	19453.26	1.07	0.28
CLM_FREQ5	20315.04	19738.57	1.03	0.30
MVR_PTS	101.51	71.01	1.43	0.15
CAR_AGE	-103.61	43.85	-2.36	0.02
PARENT1Yes	30.94	676.44	0.05	0.96
SEXF	-1138.26	608.03	-1.87	0.06
EDUCATIONBachelors	375.91	651.75	0.58	0.56
EDUCATIONMasters	1422.22	1103.62	1.29	0.20
EDUCATIONPhD	2862.22	1337.73	2.14	0.03
EDUCATIONz_High School	-394.64	517.11	-0.76	0.45
JOBClerical	246.82	1209.01	0.20	0.84
JOBDoctor	-2551.55	1776.31	-1.44	0.15
JOBHome Maker	-175.69	1309.93	-0.13	0.89
JOBLawyer	226.52	1035.14	0.22	0.83
JOBManager	-871.00	1070.19	-0.81	0.42
JOBProfessional	1060.80	1132.34	0.94	0.35
JOBStudent	-36.23	1316.43	-0.03	0.98
JOBz_Blue Collar	526.53	1150.84	0.46	0.65
CAR_TYPEPanel Truck	-141.28	1000.97	-0.14	0.89
CAR_TYPEPickup	-70.84	600.81	-0.12	0.91
CAR_TYPESports Car	957.50	757.21	1.26	0.21
CAR_TYPEVan	133.26	773.88	0.17	0.86
CAR_TYPESUV	818.76	676.55	1.21	0.23
REVOKEDYes	-1235.84	531.61	-2.32	0.02
URBANICITYRural	-62.29	764.37	-0.08	0.94
MSTATUSNo	697.44	469.32	1.49	0.14
CAR_USEPrivate	-438.54	525.54	-0.83	0.40

Standard errors: OLS

Observations	8161
Dependent variable	TARGET_AMT
Type	OLS linear regression

F(51,8109)	12.55
R^2	0.07
Adj. R^2	0.07

	Est.	S.E.	t val.	p
(Intercept)	-2940.11	6006.37	-0.49	0.62
KIDSDRIV1	607.79	212.58	2.86	0.00
KIDSDRIV2	585.13	303.08	1.93	0.05
KIDSDRIV3	518.28	602.63	0.86	0.39
KIDSDRIV4	-134.41	2377.30	-0.06	0.95
log(AGE)	-39.22	316.19	-0.12	0.90
HOMEKIDS1	140.49	209.15	0.67	0.50
HOMEKIDS2	225.40	204.47	1.10	0.27
HOMEKIDS3	60.28	240.39	0.25	0.80
HOMEKIDS4	72.16	396.93	0.18	0.86
HOMEKIDS5	437.67	1276.10	0.34	0.73
YOJ	18.22	18.84	0.97	0.33
log(INCOME + 0.00000000000001)	-13.00	7.49	-1.74	0.08
INCOME	-0.00	0.00	-2.18	0.03
HOME_VAL	-0.00	0.00	-1.16	0.25
log(TRAVTIME)	191.61	221.88	0.86	0.39
TRAVTIME	5.09	8.50	0.60	0.55
log(BLUEBOOK)	97.20	212.87	0.46	0.65
BLUEBOOK	0.01	0.02	0.53	0.59
TIF	-47.60	12.20	-3.90	0.00
log(OLDCLAIM + 0.00000000000001)	-91.75	172.76	-0.53	0.60
OLDCLAIM	-0.01	0.02	-0.57	0.57
CLM_FREQ1	4155.17	6950.73	0.60	0.55
CLM_FREQ2	4035.61	6945.68	0.58	0.56
CLM_FREQ3	4219.65	6948.61	0.61	0.54
CLM_FREQ4	4224.96	6963.30	0.61	0.54
CLM_FREQ5	3942.12	7050.82	0.56	0.58
MVR_PTS	163.67	26.76	6.12	0.00
CAR_AGE	-31.91	12.77	-2.50	0.01
PARENT1Yes	461.26	217.58	2.12	0.03
SEXF	-352.56	162.52	-2.17	0.03
EDUCATIONBachelors	-206.58	206.51	-1.00	0.32
EDUCATIONMasters	125.82	303.83	0.41	0.68
EDUCATIONPhD	421.14	361.74	1.16	0.24
EDUCATIONz_High School	-66.64	172.21	-0.39	0.70
JOBClerical	480.68	342.31	1.40	0.16
JOBDoctor	-521.54	409.01	-1.28	0.20
JOBHome Maker	184.60	375.30	0.49	0.62
JOBLawyer	216.11	295.78	0.73	0.47
JOBManager	-472.63	288.46	-1.64	0.10
JOBProfessional	445.55	308.89	1.44	0.15
JOBStudent	109.87	384.54	0.29	0.78
JOBz_Blue Collar	491.49	322.31	1.52	0.13
CAR_TYPEPanel Truck	266.97	283.04	0.94	0.35
CAR_TYPEPickup	381.46	170.82	2.23	0.03
CAR_TYPESports Car	1033.80	217.41	4.76	0.00
CAR_TYPEVan	498.21	213.21	2.34	0.02
CAR_TYPESUV	754.27	179.60	4.20	0.00
REVOKEDYes	593.79	175.58	3.38	0.00
URBANICITYRural	-1624.11	141.06	-11.51	0.00
MSTATUSNo	599.55	152.00	3.94	0.00
CAR_USEPrivate	-772.64	164.49	-4.70	0.00

Standard errors: OLS